Explainable AI Methods - A Brief Overview

Year: 2020 | Citations: 321 | Authors: Andreas Holzinger, Anna Saranti, Christoph Molnar, P. Biecek, W. Samek

Abstract

. Explainable Artificial Intelligence (xAI) is an established field with a vibrant community that has developed a variety of very successful approaches to explain and interpret predictions of complex machine learning models such as deep neural networks. In this article, we briefly introduce a few selected methods and discuss them in a short, clear and concise way. The goal of this article is to give beginners, especially application engineers and data scientists, a quick overview of the state of the art in this current topic. The following 17 methods are covered in this chapter: LIME, Anchors, GraphLIME, LRP, DTD, PDA, TCAV, XGNN, SHAP, ASV, Break-Down, Shapley Flow, Textual Explanations of Visual Models, Integrated Gradients, Causal Models, Mean-ingful Perturbations, and X-NeSyL.